Brier Score

Another common validation metric is the Brier score. Here we have the same setting as before, but our outcome Y is binary. The lower the Brier score is for a set of predictions, the better the predictions are calibrated. Note that the Brier score, in its most common formulation, takes on a value between zero and one, since this is the largest possible difference between a predicted probability (which must be between zero and one) and the actual outcome (which can take on values of only 0 and 1).

Setting

- $(X, Y) \sim \wp$, where:
  - Y is a binary random variable
  - X is a random vector
  - $\wp$ is their unknown joint distribution
- M(X) prediction model: X $\mapsto$ [Prediction of Y]
- Dataset:

$$\begin{pmatrix} X_1 & Y_1 \\ . & . \\ . & . \\ . & . \\ X_N & Y_N \end{pmatrix}$$

Example: Logistic Regression

$$Y_i = \frac{e^{\beta_0 + \beta_1 X_i}}{1 + e^{\beta_0 + \beta_1 X_i}}$$

$$M(X_i) = \frac{e^{\widehat{\beta_0} + \widehat{\beta_1} X_i}}{1 + e^{\widehat{\beta_0} + \widehat{\beta_1} X_i}}$$

- $Z(M(X), \wp) = \mathbb{E}[(M(X) - Y)^2]$ w.r.t. $\wp$
- How to estimate Z?

$$\frac{1}{N} \sum_{i=1}^{N} (M(X_i) - Y_i)^2$$

Example

Upload the data set `birthwt` from the library MASS in R by the commands:

```
library(MASS)
data(birthwt)
attach(birthwt)
str(birthwt)
```

```
## 'data.frame':    189 obs. of  10 variables:
##  $ low  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age  : int  19 33 20 21 18 21 22 17 29 26 ...
##  $ lwt  : int  182 155 105 108 107 124 118 103 123 113 ...
##  $ race : int  2 3 1 1 1 3 1 3 1 1 ...
##  $ smoke: int  0 0 1 1 1 0 0 0 1 1 ...
##  $ ptl  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ht   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ui   : int  1 0 0 1 1 0 0 0 0 0 ...
##  $ ftv  : int  0 3 1 2 0 0 1 1 1 0 ...
##  $ bwt  : int  2523 2551 2557 2594 2600 2622 2637 2637 2663 2665 ...
```

The dataset has 189 observations and 10 variables. It contains the data from a study that was conducted to identify the risk factors on low birth weight by Baystate Medical Center, Springfield, in Massachusetts during 1986. The variable of interest is the binary variable `low`, which takes value 1 if the birth weight was less than 2.5 kg and 0 otherwise.

The logistic regression is fitted by the command `glm()`, where we need to specify the `family` (e.g. binomial) and the `link function` (e.g logit). Let's try several logistic models and look at their output:

```
race = factor(race) # Let's treat this variable as categorical

# Model 1
mod1 = glm(low ~ lwt + race + age + ftv, family = binomial(link = logit))
summary(mod1)
```

```
##
## Call:
## glm(formula = low ~ lwt + race + age + ftv, family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.4163  -0.8931  -0.7113   1.2454   2.0755
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.295366   1.071443   1.209   0.2267
## lwt         -0.014245   0.006541  -2.178   0.0294 *
## race2        1.003898   0.497859   2.016   0.0438 *
## race3        0.433108   0.362240   1.196   0.2318
## age         -0.023823   0.033730  -0.706   0.4800
## ftv         -0.049308   0.167239  -0.295   0.7681
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
```

```
## Residual deviance: 222.57  on 183  degrees of freedom
## AIC: 234.57
##
## Number of Fisher Scoring iterations: 4
```

```r
# Model 2
mod2 = glm(low ~ lwt + race, family = binomial(link = logit))
summary(mod2)
```

```
##
## Call:
## glm(formula = low ~ lwt + race, family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.3491  -0.8919  -0.7196   1.2526   2.0993
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.805753   0.845167   0.953   0.3404
## lwt         -0.015223   0.006439  -2.364   0.0181 *
## race2        1.081066   0.488052   2.215   0.0268 *
## race3        0.480603   0.356674   1.347   0.1778
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 234.67  on 188  degrees of freedom
## Residual deviance: 223.26  on 185  degrees of freedom
## AIC: 231.26
##
## Number of Fisher Scoring iterations: 4
```

The corresponding Brier scores are obtained by hand as

```r
# Model 1
pred1 = predict(mod1, type='response') # type='response' gives the predicted probabilities
brier1 = mean((pred1 - low)^2)
print(brier1)
```

```
## [1] 0.2018814
```

```r
# Model 2
pred2  = predict(mod2, type='response')
brier2 = mean((pred2 - low)^2)
print(brier2)
```

```
## [1] 0.2022583
```

The Brier Score can be easily obtained also through the function `verify()` in the library verification:

```
library(verification)
verify(low, pred1)$bs
```

```
## If baseline is not included, baseline values  will be calculated from the  sample obs.
```

```
## [1] 0.2007011
```

```
verify(low, pred2)$bs
```

```
## If baseline is not included, baseline values  will be calculated from the  sample obs.
```

```
## [1] 0.2030291
```